Tipping pro-environmental norm diffusion at scale: opportunities and limitations

Joël Berger [1, 2], Charles Efferson [3], Sonja Vogt [3, 4]

[1] Department of Business and Management, Kalaidos University of Applied Sciences, Switzerland

[2] Institute of Sociology, University of Bern, Switzerland

[3] Faculty of Business and Economics, University of Lausanne, Switzerland

[4] Nuffield College, University of Oxford, UK

Correspondence:
joel.berger@kalaidos-fh.ch; charles.efferson@unil.ch; sonja.vogt@unil.ch; Joël Berger, Kalaidos University of Applied Sciences, Department of Business and Management, Jungholzstrasse 43, 8050 Zürich, +41 44 200 19 04

A revised version of this manuscript has been accepted for publication in Behavioural Public Policy.

Cite as: Berger, Joël; Efferson, Charles; Vogt, Sonja. 2021. Tipping pro-environmental norm diffusion at scale – Opportunities and limitations. Behavioural Public Policy, 1-26, doi:10.1017/bpp.2021.36

Abstract: Rapid and comprehensive social change is required to mitigate pressing environmental issues such as climate change. Social tipping interventions have been proposed as a policy tool for creating this kind of change. Social tipping means that a small minority committed to a target behaviour can create a self-reinforcing dynamic, which establishes the target behaviour as a social norm. The possibility of achieving the large-scale diffusion of pro-environmental norms and related behaviours with an intervention delimited in size and time is tempting. Yet the canonical model of tipping, the coordination game, may evoke overly optimistic expectations regarding the potential of tipping, due to the underlying assumption of homogenous preferences. Relaxing this assumption, we devise a threshold model of tipping pro-environmental norm diffusion. The model suggests that depending on the distribution of social preferences in a population, and the individual cost of adopting a given pro-environmental behaviour, the same intervention can activate tipping, have little effect, or produce a backlash. Favourable to tip pro-environmental norms are widespread advantageous inequity aversion and low adoption costs. Adverse are widespread self-regarding preferences or disadvantageous inequity aversion, and high costs. We discuss the policy implications of these findings and suggest suitable intervention strategies for different contexts.

Keywords: coordination, pro-environmental behaviour, norm change, norm diffusion, social tipping, threshold

Introduction

Behavioural science interventions for modifying human behaviour, in the following called (behaviour) interventions, have been put forward to mitigate large-scale environmental problems, such as climate change or the ongoing loss of biodiversity (Amel *et al.*, 2017; Reddy *et al.*, 2017; Battista *et al.*, 2018; Klotz *et al.*, 2018; Travers *et al.*, 2021). These interventions typically rely on human decision heuristics or social influence (Byerly *et al.*, 2018; Cinner, 2018). For example, default nudging exploits the heuristic of simply going with the default option instead of pondering which option is best (Thaler and Sunstein, 2008). As an application, providers have set sustainably produced electricity instead of conventionally produced electricity as a default (Liebe, Gewinner and Diekmann, 2021). Interventions based on social influence typically stress behaviours that are socially expected or chosen by a majority. For example, to encourage the re-usage of towels in hotels, guests have been made aware that a majority is engaging in this behaviour (Goldstein, Cialdini and Griskevicius, 2008).

A number of quantitative meta-analyses show that sometimes these kinds of interventions can have a tremendous impact. Yet, the typical effect size is moderate at best (Osbaldiston and Schott, 2012; Abrahamse and Steg, 2013; Lokhorst *et al.*, 2013; Bergquist, Nilsson and Schultz, 2019; Nisa *et al.*, 2019), with a tendency for methodologically more rigorous studies to report smaller effects (Delmas, Fischlein and Asensio, 2013; Nisa *et al.*, 2019).

Here, we discuss a number of reasons why interventions have mixed effects, which will in turn point toward specific strategies for improvement. Our starting point is the observation that most interventions target individuals to trigger behaviour change (Bicchieri and Dimant, 2019). Models of cultural evolution suggest targeting groups instead because a strategy of this kind can reinforce the impact of a given intervention (Waring *et al.*, 2015; Waring, Goff and Smaldino, 2017; Brooks *et al.*, 2018).

Pioneering theoretical contributions have provided fascinating insights into the mechanisms that translate behaviour change of a few into large-scale societal change. These contributions centre on the notion of 'social tipping'. Tipping is the phenomenon that a small subgroup of a population, adopting a new behaviour, can trigger a self-reinforcing diffusion process that finally establishes this behaviour as a social norm – provided the size of the corresponding subgroup crosses a critical value called the 'tipping point' (Milkoreit *et al.*, 2018; Nyborg, 2020; Scheffer, 2020). Here, we define a social norm in a descriptive sense, as a behavioural pattern that individuals prefer to conform to when they believe that a sufficient number of others is also conforming to this pattern (Bicchieri, 2005).

Recent research empirically corroborates the power of tipping. Subgroups of 25-35 per cent of the population can establish a new behaviour as a social norm in the laboratory (Centola *et al.*, 2018; Andreoni, Nikiforakis and Siegenthaler, 2021). Concerning social norms regulating behaviours with environmental impact (in the following called pro-environmental norms, and pro-environmental behaviours), simulation models come to similar results. Behaviour change of a minority can establish,

but also erode, pro-environmental norms (Walker and Meyers, 2004; Castilla-Rho *et al.*, 2017; Sigdel, Anand and Bauch, 2017; Mathias *et al.*, 2020) – a theoretical result in line with the findings of a recent field study (Berger, 2021).

Tipping has been applied in attempts to resolve various social issues (Mackie and Lejeune, 2009; Organisation, 2009; Shell-Duncan *et al.*, 2011; Dolan *et al.*, 2012; Lee-Rife *et al.*, 2012; World Bank Group, 2015; Nyborg *et al.*, 2016; Bicchieri, 2017; O'Brien, 2017; Platteau and Auriol, 2018). The idea of recruiting the endogenous dynamics of norm diffusion with an exogenous intervention, delimited in size and time, is tempting and has recently been put forward as a tool for environmental policy (Westley *et al.*, 2011; Nyborg, 2018, 2020; Otto *et al.*, 2020). Yet there is no foolproof way of activating tipping, and attempts to do so have produced a variety of results (Bicchieri, 2017; Efferson, Vogt and Fehr, 2020).

Tipping dynamics have so far been described with coordination game models (Schelling, 1973; Bicchieri, 2005; Nyborg, 2020). We first outline how these models apply to large-scale pro-environmental behaviour change. Then, we show that a simplifying assumption of these models, the assumption of homogenous preferences, may evoke overly optimistic expectations about the potential of tipping. More to the point, recent research in a related field, the abandonment of harmful traditions (Efferson, Vogt and Fehr, 2020; Efferson, 2021), shows that preference heterogeneity can constrain tipping.

Here, we discuss the consequences of preference heterogeneity for the potential of tipping pro-environmental norm diffusion. In particular, analysing a threshold model of

collective behaviour (Granovetter, 1978; Schelling, 1978; Macy, 1991), we address the following questions. Which structures of preference heterogeneity enable or constrain tipping? And, given a specific structure, which intervention strategy is the most promising? Our research holds clear implications for environmental policy.

Tipping pro-environmental norm diffusion

Coordination incentives and social tipping

Many environmental problems can theoretically be viewed as a social dilemma, for example, a public goods game (Milinski, Semmann and Krambeck, 2002; Irwin, 2009; Boyd *et al.*, 2012; Hauser *et al.*, 2016; Goeschl *et al.*, 2020). In public goods games, public and private benefit are in conflict. Everyone profits from environmental public goods such as a stable climate. At the same time, everyone has an individual incentive to withhold their contribution to environmental public goods, for example by taking convenient short-haul flights instead of time-consuming train rides (Gössling, Humpe and Bausch, 2020). Under the traditional assumption of self-interested preferences, an implication of this model is the tragedy of the commons (Hardin, 1968).

Yet, a common only ends tragically when it is not governed properly. Numerous environmental dilemmas around the world have been solved successfully – with the help of institutions (Ostrom, 2015). Implementing a formal institution, however, is not always feasible. Informal institutions, in turn, typically rely on the interaction of familiar individuals and are therefore not well-suited for anonymous large-scale societies (Guala, 2012).

One exception are social norms. Norms regulate behaviour not only in small groups but also at scale (Mackie, 1996; Nyborg and Rege, 2003; Young, 2015) – a general finding,

that extends to pro-environmental behaviour (Walker and Meyers, 2004; Nyborg *et al.*, 2016; Castilla-Rho *et al.*, 2017; Sigdel, Anand and Bauch, 2017). For this reason, the diffusion of norms has been suggested for regulating behaviours with global environmental impact (Boyd *et al.*, 2012; Nyborg *et al.*, 2016; Nyborg, 2018, 2020; Otto *et al.*, 2020).

Since Hardin, a large body of literature has shown that many people are not only motivated by self-interest but also by social preferences. Social preferences, as we will show, enable the diffusion of pro-environmental norms. In particular, a specific form of social preference, conditional cooperation, is a major driver of cooperation in experimental public goods games (Henrich *et al.*, 2001; Fehr and Fischbacher, 2003; Gintis *et al.*, 2003). While self-interested individuals free-ride in these games, conditional cooperators are willing to cooperate to the extent that others are cooperating, too (Fischbacher, Gächter and Fehr, 2001; Fehr and Gächter, 2002; Biel and Thøgersen, 2007; Chaudhuri, 2011) – a finding that extends to real-world environmental dilemmas (Rustagi, Stefanie and Kosfeld, 2010; Fehr and Leibbrandt, 2011; Bluffstone *et al.*, 2020; Goeschl *et al.*, 2020).

In a social dilemma, mutual defection is the only Nash equilibrium. No matter how many others are travelling by train, as a strategy of cooperation (C), instead of taking more convenient short-haul flights, as a strategy of defection (D), a purely self-interested actor will always defect (Goeschl *et al.*, 2020). It is important to note that beliefs about the probability for others to choose C are therefore irrelevant. Cooperation is unconditionally disfavoured (for more details refer to Figure A1, panels a and b, in the Appendix).

What changes under the empirically more accurate assumption that some people are conditionally cooperative? We elaborate on this question using Fehr and Schmidt's (1999) influential theory of social preferences. This theory extends the standard model with the assumption that some people dislike inequity. More specifically, they dislike being worse off than others – a preference called disadvantageous inequity aversion, α. And they dislike being better off than others, a preference called advantageous inequity aversion, β. Assume two conditional cooperators, 'focal' and 'other', with identical Fehr-Schmidt preferences, playing a nominal social dilemma game with two options, cooperation and defection (Figure A1c in the Appendix). Importantly, for an actor with associated preferences, not only the nominal payoffs are relevant. Also disadvantageous and advantageous inequality matter. For this reason, unlike a self-interested actor, a conditional cooperator will not simply defect. Instead, she prefers to cooperate, provided other is cooperating, too. A conditional cooperator is willing to endure a time-consuming train ride when many others are also doing so.

More technically, the greater a conditionally cooperative focal's expected probability, q, that other is taking the train, the greater her expected utility from cooperation, $E(C)$. Equating $E(C)$ with her expected utility of defection, $E(D)$, yields focal's indifference point, q^*. At this point, she is indifferent between C and D (Figure A1d in the Appendix). After crossing q^*, focal prefers C over D which means, the nominal social dilemma turns into a coordination game at this point (Bicchieri, 2005; Bowles, 2009; Centola, 2013).

In a coordination game, mutual defection is no longer the only Nash equilibrium. Instead, there are two equilibria: mutual cooperation and mutual defection. When the game is played repeatedly in a group of conditional cooperators, the population dynamics can converge to either of these equilibria. There is a third equilibrium, namely one in mixed strategies. That is, in absence of a best strategy, people could take a random choice – they could choose C with probability q^* and D with probability $1 - q^*$ (Sugden, Harsanyi and Selten, 1989). However, the equilibrium in mixed strategies, q^*, is unstable. Even a small deviation of the population dynamics from q^* towards either mutual cooperation or mutual defection activates a self-reinforcing shift towards one of the pure strategy equilibria. For this reason, q^* is also called a 'tipping point' (Nyborg, 2020).

As an illustration, conditional cooperators choose 8-hour train rides over 2-hour flights when they believe that many others are doing so, too. That is, each time before travelling, they update their beliefs about the current proportion of train users, q_t. When $q_t < q^*$, they opt for flights and the proportion of train users decreases over time, converging towards zero. When $q_t > q^*$, they opt for train travel, and the proportion of train users increases, converging towards the cooperative equilibrium.

Let us now assume a policymaker intending to push the population dynamics towards the cooperative equilibrium to curb climate gas emissions. Most policymakers have limited resources, and they are therefore unable to target the entire population with an intervention. Luckily, in the presence of coordination incentives, there is a more efficient option. Namely the option of treating only a delimited sample, the target, in an attempt to push the population over the tipping point. When successful, this would then

trigger an endogenous process of norm diffusion. More specifically, such an intervention has two hypothetical effects. A direct effect, enhancing climate-friendly choices in the target. And an indirect effect in the untreated segment of the population, emulating the behaviour of the treated. We call this indirect effect 'social tipping', or simply 'tipping'. In our example, the policymaker could deliberately try to create tipping by pushing the proportion of train users in the population across the critical value, q^*, with a delimited intervention.

When the proportion of train users, q_t, does not undercut q^* too strongly, pre-intervention, the policymaker might well succeed. As soon as q^* is crossed, more and more individuals switch from plane to train – an endogenous process of change then successively establishes train travel as a social norm, followed by everyone. In contrast, when the proportion of train users substantially undercuts q^*, even a strong intervention might fail to push train travel across the tipping point, q^* (Nyborg, 2020). In other words, for a policymaker to succeed, it is crucial to know q^*. The higher q^*, the harder the cooperative equilibrium is to reach.

It can be shown that q^* depends on two parameters – the incentive structure of the nominal game and, the strength of social preferences relative to self-interest in the (homogenous) population (see Appendix, note on Figure A1d.).

Concerning incentives, the higher the tension between public and private benefit, the higher the threshold q^*. An example of a high-tension situation is the choice between air travel and train travel. An individual, enduring an 8-hour train ride instead of taking a

2-hour flight, bears relatively high costs in the form of effort, time and money, while the reduction in emissions achieved is very limited at the global scale. An example of a low-tension situation is switching off the lights when leaving a public bathroom, which also has a limited impact, but requires little effort.

This model implication fits well the following results of recent meta-analyses: changing high-impact behaviours, such as avoiding flights or animal products (Wynes and Nicholas, 2017; Schiermeier, 2019), seems difficult to achieve with behaviour interventions, compared to behaviours with relatively low impact (Bergquist, Nilsson and Schultz, 2019; Nisa *et al.*, 2019, see Figure 1). As a likely explanation, changing high-impact behaviours typically requires individually costly lifestyle changes. Put differently, for these behaviours, the tension between private and public benefit in the underlying environmental dilemma is quite strong.

Granted, Figure 1 provides only suggestive evidence for the claim that tension is a key factor of intervention effectiveness. One potential confounder is the type of intervention strategy. Specific research paradigms, favouring strategies that vary with respect to their strength or impact, tend to address specific behaviours (Nisa *et al.*, 2019), which could at least partly explain the association between tension and effectiveness.

In defence of our claim that tension largely controls effectiveness, consider a robust result from environmental sociology. Namely, the result that pro-environmental attitudes predict quite well pro-environmental behaviours that require little individual cost or effort, compared to a non-environmentally friendly standard. In contrast, attitudes

are much less predictive of costly pro-environmental behaviours (Diekmann and Preisendörfer, 2003; Steg *et al.*, 2014; Moser and Kleinhückelkotten, 2018; Farjam, Nikolaychuk and Bravo, 2019).

========== Figure 1 about here ==========

Concerning social preferences, one can show that for a given environmental dilemma, in a homogenous population with identical Fehr-Schmidt preferences, the following holds. The stronger advantageous inequity aversion β relative to disadvantageous inequity aversion α and self-interest, the lower the perceived tension between public and private interest, and therefore, the lower q^*. This implies that two environmental dilemmas with the same incentive structure could be resolved in one population, where advantageous inequity aversion dominates self-interest, and remain unsolved in another population, where the opposite holds. This finding is crucial, as social preferences vary substantially between societies (Henrich et al., 2010, 2006).

Introducing heterogeneity – a threshold model of social tipping

Not only do social preferences vary across societies. Within society, they also vary across individuals (Engelmann and Strobel, 2004; Blanco, Engelmann and Normann, 2011). In any human population, we will not observe α and β values common to everyone, but distributions of α and β values. As a consequence, we will not observe a common threshold, q^*, but a distribution of individual thresholds, q_i^*, as suggested by threshold models of collective behaviour (Granovetter, 1978; Schelling, 1978; Macy, 1991).

These models propose that, once the proportion of cooperators in the population reaches an individual's threshold, q_i^*, this specific individual switches from defection to cooperation. But individuals vary concerning the value that makes them switch. It is a crucial model implication that the structure of threshold heterogeneity in a population determines the potential for tipping. Efferson et al. (2020) have shown for abandoning harmful traditions that threshold heterogeneity still plays the dominant role when adding a number of complexities, for example, different network structures or intervention strategies.

In what follows, based on a threshold model, we discuss how preference heterogeneity and the tension inherent to a specific environmental dilemma control the potential for tipping the diffusion of pro-environmental norms. Consider Online Supplementary Information 1 for a more technical discussion of the model.

Let F be a cumulative distribution function of individual thresholds, q_i^*, and q_t the proportion of cooperators at time t. When each individual in a population regularly evaluates the cooperative behaviour of others and updates her beliefs accordingly, then cooperation changes over time according to $q_{t+1} = F(q_t)$. This is a well-discussed property of the threshold model (Granovetter, 1978; Schelling, 1978). As a central implication, depending on the shape of F, a social tipping point may exist, but not necessarily. In other words, the presence of coordination incentives does not necessarily imply a strong potential for tipping. Depending on the shape of F, the potential for tipping can vary tremendously.

More specifically, the tension inherent to a social dilemma and the distribution of social preferences in the population together control the shape of a threshold distribution, with strong implications for tipping.

Under right-skewed threshold distributions, the potential for tipping is considerable. In this scenario, a majority has low thresholds. Some may even be willing to cooperate unconditionally, and only a minority has high thresholds. This kind of distribution results from an environmental dilemma with low inherent tension, such as switching off the lights when leaving a public bathroom – especially, when in combination with widespread advantageous inequity aversion. Figure 2a shows an illustrative right-skewed threshold distribution, and Figure 2d shows the corresponding cumulative distribution. Points at which F intersects the diagonal from above are stable equilibria, whereas points at which F intersects from below are unstable equilibria. Here, F intersects from above at $F(0) = 0$ and $F(1) = 1$, implying two stable equilibria, zero cooperation

and total cooperation. And it intersects from below, at $F(0.1) = 0.1$, which suggests that a tipping point exists and that this tipping point is low. Put differently, should a policymaker convince slightly more than 10% of the population to switch off the lights regularly, this behaviour would then spread without further ado until adopted by 100% of the population.

Under left-skewed distributions, the potential for tipping is the smallest. In this scenario, a majority has high thresholds, some may even defect unconditionally, and only a small minority of very prosocial individuals has low thresholds. An example is avoiding flights. As can be seen from the corresponding cumulative distribution in Figure 2f, a tipping point still exists in this example, but with $F(0.9) = 0.9$, it may be difficult to reach, depending on the current level of cooperation in the population. Moreover, the potential for tipping only amounts to 10% of the population.

Symmetrical distributions are an intermediate case. They can result from environmental dilemmas with intermediate inherent tension, such as bringing one's empty glass bottles to public recycling stations, in combination with moderate levels of social preferences. In Figure 2b, a tipping point lies at $F(0.5) = 0.5$, and the potential for tipping amounts to 50 percentage points (Figure 2e). Obviously the potential can still be substantial under symmetric distributions, and, depending on the resources invested, a tipping process could still be activated.

In brief, any intervention is most likely to activate tipping under right-skewed distributions, and least likely under left-skewed distributions, with symmetrical distributions as an intermediate case.

========== Figure 2 about here ==========

It is worth noting, however, that, unlike in the three examples discussed, a tipping point might not even exist. For example, a threshold distribution could be even more left-skewed than the one shown in 2c, with substantial population shares defecting unconditionally. In this case, the corresponding cumulative distribution stays below the 45°-line entirely, pointing to the absence of a tipping point. As an additional possibility, distinct groups in a threshold distribution may constrain tipping. Picture two opposite groups in a population, one amenable and one opposed to climate change mitigation, as currently the case in the US (Mccright and Dunlap, 2011; Kahan, 2012, 2017; Rinscheid, Pianta and Weber, 2020). Tipping the diffusion of climate protecting behaviours among Democrats may be quite feasible, but these behaviours are unlikely to spill over from Democrats to Republicans under pronounced political polarization. In this case, cutting

the link between climate change mitigation and Republican identity is a pre-condition for tipping (Doell *et al.*, 2021).

That said, under the threshold model, a policymaker intending to apply tipping to environmental policy, has three options. She can treat the beliefs regarding the current cooperation level in the population, leaving the thresholds unchanged (1). She can target the thresholds by changing the monetary incentives of the nominal game in favour of cooperation (2). Or she can change the thresholds by changing the underlying preferences – a strategy that strengthens the intrinsic value of cooperation (3).

Changing beliefs

The first strategy, changing beliefs about the prevalence of the target behaviour, has frequently been used to promote pro-environmental behaviour, under the label 'descriptive norms intervention' (e.g. Schultz *et al.*, 2007; Goldstein, Cialdini and Griskevicius, 2008). The key idea is feeding back the current cooperation level, called descriptive information, to the population, which then leads people to adapt false beliefs. The behavioural consequence of the intervention, which makes the distribution of choices at any point in time public knowledge, though, is not straightforward. Beliefs-based interventions can produce a variety of outcomes, including strong beneficial effects, nil results, and backlash (Schultz *et al.*, 2007; Allcott and Rogers, 2014; Farrow, Grolleau and Ibanez, 2017; Rinscheid, Pianta and Weber, 2020). For example, feeding back the average energy consumption of a neighbourhood to the households of this neighbourhood

promoted energy saving among households with above-average consumption, but increased consumption among households below the average (Schultz *et al.*, 2007). In principle, due to the no-control group design of that specific study, regression to the mean could explain this result. That is, the 'backlash' among households with below-average energy consumption might have occurred even in absence of the intervention (Verkooijen, Stok and Mollen, 2015). Yet regression to the mean it is an unlikely explanation for the backlash observed in many other studies. For example, providing feedback about the share of coffee sold in reusable mugs instead of one-way paper cups increased the share of coffee bought in mugs at one cafeteria, with a relatively high initial level of mug usage, compared to a control group. At the same time, the intervention discouraged mug usage at another cafeteria, with a relatively low initial level of mug usage, compared to the same control group (Berger, 2021). In fact, evidence from controlled laboratory experiments suggests that an individual's beliefs about the prevalence of cooperation in a group are a strong determinant of that individual's cooperativeness (Fischbacher, Gächter and Fehr, 2001; Bicchieri and Xiao, 2009; Ackermann and Murphy, 2019). Based on this result, the threshold model has a clear answer to the question of when beliefs-based interventions have beneficial outcomes, and when harmful outcomes are more likely.

To illustrate this, picture a policymaker intending to tip the diffusion of a pro-environmental norm, with the provision of regular feedback about the current cooperation level to the population. The standard version of the model assumes actors that regularly update their beliefs regarding the prevalence of cooperation, and it assumes

these beliefs to be accurate. In the real world, the latter assumption may be violated. In fact, in the absence of clear information about the prevalence of a specific pro-environmental behaviour, people tend to underestimate the engagement of others, relative to their own engagement – a bias that dampens their willingness to act environmentally friendly (Pieters *et al.*, 1998; Bergquist, 2020; Leviston and Uren, 2020). In the following discussion, we therefore explicitly assume beliefs that are distorted and that can also be coupled to preferences, pre-intervention. We further assume that the individuals targeted by an intervention adjust their beliefs to the descriptive information provided. Thus, belief formation after intervention is based on accurate information, but it is not forward looking. Instead, it is myopic, and people best respond given these beliefs, as in the original threshold model (Granovetter, 1978). We discuss an intervention that provides descriptive feedback regularly. Regular feedback provision promotes the regular updating of beliefs, as assumed by the model. For now, we also assume that the entire population is treated by the intervention and discuss the strategy of treating only a delimited sample later. Online Supplementary Information 2 outlines a threshold model with distorted beliefs and shows how our key findings derive from the model.

How does the joint distribution of thresholds and beliefs shape the net effect of a beliefs-based intervention, under these assumptions? As a starting point, consider Figure 3a, showing 500 thresholds, q_i^*, and beliefs, $\hat{q}_{it}$, pre-intervention. Individuals above the 45°-line hold beliefs matching or exceeding their thresholds, they therefore cooperate. Individuals below the 45°-line hold beliefs smaller than their thresholds, they therefore defect.

As we will demonstrate, the long-term effect of an intervention depends on the relative size of four subgroups, pre-intervention. First, cooperators with accurate beliefs in the sense that they believe cooperation to match or exceed their thresholds, which is true (x_1 in region A of Figure 3a). It is worth noting that actual cooperation may overshoot or undershoot their beliefs to some degree, which is, however, inconsequential. What matters is that these agents presume cooperation to match or exceed their thresholds and that this is actually the case. Cooperators holding beliefs that are accurate in this sense cooperate before and after the first round of feedback. Second, cooperators with false beliefs, in the sense that they assume cooperation to match or exceed their thresholds, while cooperation, in fact, undershoots their thresholds (x_2 in region B of Figure 3a). They cooperate before the intervention, but, disappointed from the first feedback, stop cooperating in response. Third, defectors holding the false belief of cooperation undershooting their thresholds, pre-intervention (x_3 in region C of Figure 3a). They defect before the intervention, but, positively surprised from the first feedback, switch to cooperation. Fourth, defectors holding the accurate belief of cooperation undershooting their thresholds, defecting before and after the first feedback (x_4 in region D of Figure 3a).

In short, a beliefs-based intervention has the following immediate effect. Disappointed cooperators (x_2) switch to defection, and positively surprised defectors (x_3) switch to cooperation. Individuals holding the accurate beliefs of cooperation exceeding their thresholds (cooperators, x_1) or undershooting their thresholds (defectors, x_4) are unaffected by the first round of the intervention. We designate $t = 1$ as the point in time when individuals who were cooperating or defecting specifically because of distorted

beliefs have changed their behaviours as an immediate response to the intervention, but no additional cultural evolutionary dynamics have yet occurred. This means, in effect, that $q_1 = q_0 - x_2 + x_3$. The immediate effect of ensuring correct beliefs can thus be positive, negative, or neutral. If $x_2 \leq x_3$ and $q_1 \geq q_0$, cooperation (weakly) rises. If $x_2 > x_3$ and $q_1 < q_0$, cooperation declines. Which of these scenarios holds will depend on the distribution of individuals in regions B and C in Figure 3a. The distribution of individuals in regions A and D is irrelevant in terms of the immediate effect of the intervention.

What happens after the immediate effect? To answer this question, note that the intervention represents a fundamental change in the informational setting. It eliminates the possibility of distorted beliefs by making the current rate of cooperation public knowledge at all points in time. As a result, beliefs become accurate, subject to the assumption of myopic updating, and they are always the same for everyone. Once this happens, the preference heterogeneity represented by the distribution of thresholds is the mechanism driving the evolution of cooperation.

We now illustrate the potential long-term outcomes of a beliefs-based intervention with more depth, discussing four scenarios. In Scenario 1 (Figure 3, a-c), the threshold distribution is right-skewed, with beliefs and thresholds uncorrelated. In this case, cooperators with accurate beliefs (x_1), and defectors with false beliefs (x_3), together outnumber cooperators with false beliefs (x_2), plus defectors with accurate beliefs (x_4). Consequently, the intervention has a positive net effect, with 100% cooperation as a potential long-term outcome, after several rounds of feedback.

In Scenarios 2, the threshold distribution is symmetrical, with thresholds and beliefs correlated positively (Figure 3, d-f). Here, the outcome depends on the relative size of those with thresholds undershooting cooperation, pre-intervention $(x_1 + x_3)$, compared to those with thresholds exceeding cooperation, pre-intervention $(x_2 + x_4)$. More specifically, when $x_1 + x_3 < x_2 + x_4$, pre-intervention, cooperation cannot exceed $x_1 + x_3$ in the long term. In contrast, when $x_1 + x_3 > x_2 + x_4$, pre-intervention, cooperation can increase, with the final result depending on the exact threshold distribution.

Scenario 3 also assumes a symmetrical distribution, but a negative correlation between thresholds and beliefs (Figure 3, g-i). In this case, individuals that cooperate or defect independently of the intervention dominate $(x_1 + x_4 > x_2 + x_3)$, implying a negligible net effect.

Finally, Scenario 4, assumes a left-skewed distribution, that is, most thresholds are of a rather high value (Figure 3, j-l). Further, thresholds and beliefs are uncorrelated. In this scenario, the dominance of high thresholds renders the disappointment of cooperators with false beliefs more likely than the positive surprise of defectors with false beliefs, with the consequence of decreased cooperation, post-intervention. This sparks a negative dynamic with zero cooperation as a likely long-term outcome.

========== Figure 3 about here ==========

In short, the outcome of a beliefs-based intervention depends on the joint distribution of thresholds and beliefs. The intervention eliminates the possibility of distorted beliefs. Once this happens, the preference heterogeneity represented by the distribution of thresholds is the mechanism driving the evolution of cooperation. Under right-skewed threshold distributions, this intervention strategy bears considerable potential, while it tends to provoke backlash under left-skewed threshold distributions. Under symmetric distributions, the outcome strongly depends on the structure of correlation between thresholds and beliefs, and the relative frequency of the four subgroups, $x_1, x_{2,} x_3$ and x_4. A negative correlation implies a negligible effect. A positive correlation implies a positive effect, when the following holds additionally. Cooperators with accurate beliefs and defectors with false beliefs, therefore switching to cooperation in response to the intervention, form a majority over defectors with accurate beliefs and cooperators with false beliefs, therefore switching to defection.

The variety of potential outcomes points to the necessity of information about the joint distribution of thresholds and beliefs for designing effective beliefs-based interventions. When feasible, this strategy is quite attractive, as feeding provision is likely cost-efficient, compared to other strategies.

Changing the thresholds with monetary incentives

The second strategy is treating the thresholds with monetary incentives, either with a subsidy, like a train voucher, or a tax, like a carbon tax on air travel. Both options enhance the benefit of cooperation relative to the benefit of defection, decreasing the thresholds. A possible drawback of monetary incentives is that the thresholds might return to their pre-intervention levels, once the incentives are removed. Even worse, monetary incentives could crowd out intrinsic motivation, with the result of increased thresholds, compared to pre-intervention levels (Frey and Oberholzer-Gee, 1997; Reeson and Tisdell, 2008; Rode, Gómez-Baggethun and Krause, 2015; Lapinski *et al.*, 2017). In a worst-case scenario, increased thresholds and lower consequent cooperation among the treated could then trigger the decay of cooperation in the population.

Changing the thresholds by changing preferences

In contrast, preference-based interventions, the third strategy, target individual preferences to decrease the thresholds. One variant is increasing the psychological value of cooperation. For example, the target individuals could learn about the environmental advantage of train travel compared to air travel. Another variant would promote advantageous inequity aversion.

It is worth noting that the combination of preference-based intervention and financial incentives has proven astonishingly effective in promoting pro-environmental behaviour. Moreover, adding a psychological intervention to the mere provision of

financial incentives seems to prevent motivation crowding-out (Kerr, Vardhan and Jindal, 2012; Kerr *et al.*, 2017, 2019). This finding is crucial – only a bundle of intervention strategies might be able to achieve the strong effect needed to activate tipping under a left-skewed threshold distribution.

Treating a delimited sample to activate tipping

It may often be the case that a policymaker lacks the resources to treat an entire population. In this case, treating only a delimited sample of the population, to activate diffusion of the target behaviour in the untreated segment of the population is an appealing approach. Yet policymakers should be aware that the shape of the threshold distributions in the population and the target largely control the potential for tipping.

More specifically, four parameters control that potential. First, the shape of the threshold distribution in the population and the target, namely, right-skewed, symmetric, or left-skewed. For the following discussion, we assume a randomly selected target, and consequently the threshold distribution in the target approximates the corresponding distribution in the population. Second, intervention effectiveness, d. We assume the intervention to decrease the thresholds in the target deterministically, by a value of d. An individual starts cooperating, when her post-intervention threshold, $q_i^{*\prime} = q_i^* - d$, is smaller than, or equal to, the cooperation level in the population, pre-intervention, q_0. The stronger d, the larger the proportion of actors in the target switching to cooperation in response to the intervention. Third, the population share targeted. Given an effective

intervention, the larger the target, the larger the share of cooperators in the population, post-intervention. Fourth, the level of cooperation in the population, pre-intervention, q_0. The higher q_0, the stronger the potential for tipping.

How does, under these assumptions, the shape of the threshold distribution in the population affect the potential for tipping? Any intervention has potentially two effects. First, a direct effect on the target. Second, an indirect effect, namely, the tipping process that is activated, once the size of the subgroup committed to the target behaviour crosses a critical value.

Concerning the direct effect, given an effective intervention, the share of those switching to cooperation in response to the intervention is largest under right-skewed distributions, where low thresholds prevail, followed by symmetric distributions, where intermediate values prevail, and, finally, left-skewed distributions, where large values prevail. Put differently, after the intervention, the subsample of cooperative individuals in the target is smaller under symmetric distributions than under right-skewed distributions, and smaller still under left-skewed distributions than under symmetric distributions. The smaller the cooperative subsample, post-intervention, the weaker the consequent potential for tipping.

The shape of the threshold distribution in the population also exerts an indirect effect on the potential for tipping. Under left-skewed distributions, high thresholds prevail, and therefore, the share of cooperators in the population needs to be larger to activate tipping, than under symmetric distributions. And under symmetric distributions,

the share of cooperators in the population needs to be larger than under right-skewed distributions. Per implication, under symmetric distributions, compared to right-skewed distributions, a smaller start-up group of cooperators coincides with the need for a larger start-up group. And under left-skewed distributions, compared to symmetric distributions, an even smaller start-up group coincides with the need for an even larger start-up group.

In combination, the direct and indirect effect create favourable conditions for tipping under right-skewed distributions, and unfavourable conditions under left-skewed distributions, with symmetric distributions as an intermediate case.

Who makes a good target?

So far, our policymaker was selecting her target individuals at random. Changing her sampling strategy could further boost her success, as Efferson et al., (2020) demonstrate. Given the policymaker has a very effective intervention at hand, an intervention that even changes the behaviour of individuals with high thresholds, selecting individuals with high thresholds as a target would further increase the potential for tipping. Simply put, the intervention then addresses the harder task of treating the less cooperative, leaving the easier task of treating the more cooperative to the endogenous process of tipping.

But selectively treating the less cooperative is also risky. Should the target individuals' responsiveness to the intervention be negatively correlated to their threshold

values, a substantial proportion of the target might be unwilling to adopt cooperation. A smaller start-up group of cooperators, in turn, implies a reduced potential for tipping.

In contrast, systematically sampling the more cooperative is counterproductive. First, some of the more cooperative might already be cooperating. Second, larger shares of relatively cooperative individuals in the target means smaller shares in the non-treated segment of the population, and, therefore, higher hurdles for tipping. The strategy of random sampling, selecting from among both, the more and the less cooperative, may therefore often be a reasonable compromise.

In the light of this finding, the widespread practice of using student samples, a subgroup typically more amenable to pro-environmental behaviour change than the population average, seems questionable (Nisa *et al.*, 2019). Not only might that sampling strategy evoke an overly optimistic view of the intervention under study. In practical applications, targeting mainly amenable student targets may reduce the potential for tipping in the population.

It is worth noting that the practice of using biased samples has been challenged before in other fields. More specifically, research on the foundations of human behaviour clearly shows that results from so-called 'WEIRD' samples, drawn from western, educated, industrialized, rich, and democratic societies, do not typically generalize to other societies (Henrich, Heine and Norenzayan, 2010b, 2010a). This insight underlines the necessity of using diverse samples for drawing robust conclusions.

Discussion and conclusion

Social tipping has been put forward as a policy tool to create the rapid social change needed for mitigating urgent environmental issues at a global scale, for example, climate change (Westley *et al.*, 2011; Nyborg *et al.*, 2016; Otto *et al.*, 2020). It is the idea of social tipping to activate a process of norm diffusion at the population level, by convincing only a delimited sample of the population to adopt an environmentally beneficial target behaviour. Should the population share adopting the target behaviour in response to a policy intervention reach a critical value, the tipping point, a self-reinforcing process then establishes this behaviour as a social norm without further interference (Milkoreit *et al.*, 2018; Nyborg, 2020; Scheffer, 2020).

The idea of recruiting the endogenous dynamics of norm diffusion with an exogenous intervention, delimited in size and time, is tempting. Yet tipping interventions have been applied in other fields than environmental policy with rather mixed results, ranging from tremendous success, over nil results, to harmful backlash (Dolan *et al.*, 2012; World Bank Group, 2015; Bicchieri, 2017). In response to these findings, some scholars have argued that the canonical model of tipping, the coordination game (Mackie, 1996; Nyborg, 2020), may give rise to somewhat optimistic expectations regarding the potentials of tipping (Efferson, Vogt and Fehr, 2020). More specifically, it is the assumption of homogenous preferences underlying the model that fuels optimism (Efferson, Vogt and Fehr, 2020; Berger, Vogt and Efferson, 2021; Efferson, 2021). People

vary with respect to their preferences, and this includes support for specific environmental policies.

Preference heterogeneity constrains tipping. This does not mean that tipping is a blunt sword under preference heterogeneity. Yet to devise a successful tipping intervention, understanding the implications of heterogeneity is key. Our research cultivates this kind of understanding, discussing the question: How does preference heterogeneity impact the potential of tipping in the field of pro-environmental policy? Applying a threshold model (Granovetter, 1978; Schelling, 1978; Macy, 1991) to the issue of pro-environmental norm diffusion, we discuss structures of preference heterogeneity more and less favourable for tipping, and suggest intervention strategies suitable for these structures.

The model starts with the notion of individual thresholds. A threshold is the population share a given individual requires to have adopted a specific behaviour, to also adopt this behaviour. It can be shown that two factors control the value of an individual threshold in an environmental dilemma. First, the tension between the public and private benefit inherent to a dilemma (Bowles, 2009; Goeschl *et al.*, 2020). The adoption of environmentally beneficial behaviours that are individually costly in terms of money, effort, or time, implies high thresholds. An example is avoiding flights. The adoption of less costly behaviours implies low thresholds. An example is switching off the lights when leaving a public bathroom (Diekmann and Preisendörfer, 2003; Steg *et al.*, 2014; Moser and Kleinhückelkotten, 2018; Farjam, Nikolaychuk and Bravo, 2019). Second, social preferences. Particularly, the stronger the individual tendency for advantageous inequity

aversion (Fehr and Schmidt, 1999), the weaker the perceived tension between public and private benefit, the lower the individual threshold (Bicchieri, 2005; Bowles, 2009; Centola, 2013).

Tension and social preferences together shape a threshold distribution. Low tension and widespread advantageous inequity aversion imply the predominance of low thresholds, or, a threshold distribution that is skewed to the right. High tension and widespread self-interest imply the predominance of high thresholds, or, a threshold distribution that is skewed to the left. Depending on shape, the potential of tipping varies substantially.

We started our discussion with a policymaker, intending to tip the diffusion of a target behaviour in a population (also called cooperation), by treating only a delimited sample of that population, the target. The exogenous intervention can achieve two effects. First, a direct effect on the target. Second, an indirect effect in the untreated segment of the population. Namely, the endogenous spreading of cooperation through the population that is activated, once the size of the subgroup adhering to this behaviour crosses a critical value, the tipping point.

Concerning the direct effect, the more strongly right-skewed the threshold distribution in the target, pre-intervention, the larger the share of the target switching from environmentally harmful to pro-environmental. And, the more left-skewed the threshold distribution in the target, pre-intervention, the smaller the share of the target switching to pro-environmental. The reason for this pattern is the following. Assume the

intervention to deterministically decrease the thresholds in the target by a specific, fixed value. When, post-intervention, a target individual's threshold undershoots the pre-intervention prevalence of the target behaviour in the population, this individual adopts the target behaviour. Holding the prevalence of the target behaviour in the population constant, interventions of identical effect size, therefore, induce adoption in larger shares of the target when low thresholds prevail (right-skewed distributions), compared to when intermediate thresholds prevail (symmetric distributions), and, even more so, compared to when high thresholds prevail (left-skewed distributions).

Concerning the indirect effect, or, tipping, again, right-skewed distributions provide favourable conditions, and left-skewed distributions conditions that are rather adverse, with symmetric distributions as an intermediate case. Simply put, when most people have low thresholds (right-skewed distribution), the size of the cooperative start-up group needed to cross a tipping point is smaller, compared to a population where roughly equal parts have low or high thresholds (symmetric distribution), or even a population, where high thresholds dominate (left-skewed distribution). Moreover, as discussed, symmetric, and even more so, left-skewed, distributions, constrain the effect of the exogenous intervention in the target, resulting in a smaller cooperative start-up group. In short, the larger the start-up group theoretically needed for activating tipping, the smaller the actual start-up group, resulting from the exogenous intervention.

What are strategies for activating tipping? A policymaker can either treat the thresholds. To do so, she could change the incentives of the nominal social dilemma underlying an environmental issue, for example, with a subsidy or tax. As an alternative,

she could treat the preferences of the individuals involved. One example is promoting advantageous inequity aversion, which would boost the psychological value of cooperation in the respective environmental dilemma.

Earlier research suggests that using financial incentives could crowd out intrinsic motivation (Frey and Oberholzer-Gee, 1997; Reeson and Tisdell, 2008; Rode, Gómez-Baggethun and Krause, 2015; Lapinski *et al.*, 2017), which would speak in favour of treating the preferences. More current research finds that combining financial incentives with psychological interventions is not only more effective than using any of the two strategies in isolation, but also prevents motivation crowding-out (Kerr, Vardhan and Jindal, 2012; Kerr *et al.*, 2017, 2019). Our findings encourage combining both strategies under left-skewed distributions, where strong effects are needed to activate tipping.

An alternative to treating the thresholds is treating the beliefs about the population prevalence of the target behaviour. Treating beliefs means simply feeding back descriptive information about this prevalence, pre-intervention, to the target. Information provision might often be more cost-efficient than offering financial incentives or implementing preference-based interventions at scale. While this is surely attractive, the outcome of a beliefs-based intervention, in particular, strongly depends on context (Schultz *et al.*, 2007; Goldstein, Cialdini and Griskevicius, 2008; Allcott and Rogers, 2014; Farrow, Grolleau and Ibanez, 2017; Rinscheid, Pianta and Weber, 2020; Berger, 2021). More specifically, beliefs-based interventions are most effective under right-skewed distributions, where low thresholds prevail, and most likely to produce backlash under left-skewed distributions, where high thresholds prevail. The reason is that backlash in

beliefs-based interventions is driven by disappointment cooperators. Namely, by individuals that expected cooperation to exceed their thresholds, but learn that the opposite is the case, therefore switching back from cooperation to defection in response to the intervention. Under left-skewed distributions, high thresholds prevail and disappointment is, therefore, more likely than positive surprise. The opposite holds under right-skewed distributions. The positively surprised defectors, learning that cooperation is exceeding rather than undershooting their thresholds, likely outnumber the disappointed cooperators, which results in a beneficial net outcome.

Under symmetric distributions, the net effect of a beliefs-based intervention depends on the correlation between thresholds and beliefs. A negative correlation implies a negligible effect. In this case, those with overly optimistic beliefs have also low thresholds. They stick to the target behaviour, post-intervention, when they learn that cooperation is exceeding their thresholds only slightly, rather than substantially. And those with overly pessimistic beliefs have high thresholds. They keep defecting when they learn that cooperation is indeed undershooting their thresholds, although to a smaller than exptected extent. Put differently, the beliefs of the majority are accurate in the sense that cooperation is smaller (defectors) or greater (cooperators) than their thresholds, as presumed. Descriptive feedback then lacks any behavioural consequence for the majority. In contrast, a positive correlation bears a strong potential for tipping, provided, the following holds additionally. Cooperators with accurate beliefs, and defectors with false beliefs, switching to cooperation in response to the intervention, form a majority over defectors with accurate beliefs and cooperators with false beliefs, switching to defection.

Therefore, it is encouraging that most people tend to underestimate other's pro-environmental engagement relative to their own (Bergquist, 2020; Leviston and Uren, 2020), as this finding suggests positive surprise might be a rather common reaction to descriptive feedback.

In short, a beliefs-based intervention only works in two kinds of situations. Either under left-skewed threshold distributions. Or under symmetric distributions, given a positive correlation between thresholds and beliefs, plus cooperators with accurate beliefs, and defectors with false beliefs, jointly forming the majority. Otherwise, beliefs-based interventions bring nil effects, if not backlash.

Finally, it is worth mentioning that large parts of a population can be completely unwilling to adopt certain pro-environmental behaviours – their thresholds are fixed at one hundred per cent, putting a drag on tipping. This is the case when a specific behaviour conflicts with social identities widely held in society (Mccright and Dunlap, 2011; Kahan, 2012, 2017; Rinscheid, Pianta and Weber, 2020). Cutting the link between the target behaviour and the social identities involved is then a precondition for a pro-environmental norm to spread through the entire population (Doell *et al.*, 2021).

To sum our findings, right-skewed distributions imply strong potentials for tipping. In this case, small interventions, targeting beliefs, preferences, or incentives, can produce quick and comprehensive norm diffusions. The opposite holds for left-skewed distributions. Here, large interventions, preferably bundling different strategies, for example, preference-based interventions plus incentive provision, are necessary for

activating tipping – if tipping can be activated at all. Notably, policymakers are well-advised not to use beliefs-based interventions in this kind of situation, as they then more likely bring backlash than beneficial effects. Symmetric distributions are an intermediate case. Even though larger interventions are necessary to activate tipping, the potential can still be relatively strong. Also, a beliefs-based strategy could work, although the potential here strongly depends on the joint distribution of preferences and beliefs. Refer to Figure 4 for a visual summary of our main findings.

=== Figure 4 about here ===

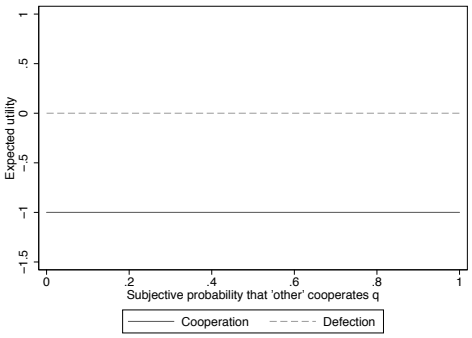
The main implication of our research is the following. A policymaker may wish to gather information about the joint distributions of thresholds and beliefs in the population, as about the population prevalence of the target behaviour, before implementing an intervention. This information will allow her to gauge the potential for tipping that is socially beneficial, and the risk of a backlash. She could then tailor her intervention to the requirements of the context. Measures of pro-environmental attitudes and behaviours, nowadays part of many large-scale surveys, may provide the information necessary. Future research should address the challenge of measuring thresholds and gauging the potential for tipping pro-environmental norm diffusion in field settings.

Appendix

a

		Other: C	Other: D
Focal	C	$b-c, b-c$	$\frac{b}{2}-c, \frac{b}{2}$
	D	$\frac{b}{2}, \frac{b}{2}-c$	$0, 0$

b



c

		Other: C	Other: D
Focal	C	$b-c, b-c$	$\frac{b}{2}-c-\alpha c,$ $\frac{b}{2}-\beta c$
	D	$\frac{b}{2}-\beta c,$ $\frac{b}{2}-c-\alpha c$	$0, 0$

d

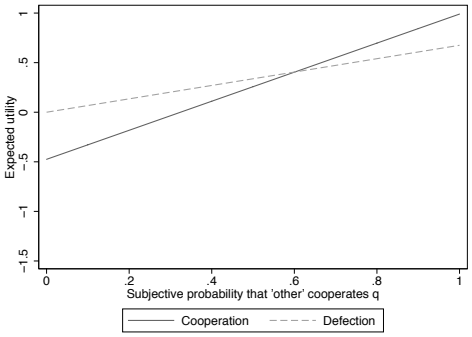


Figure A1. Public goods game and coordination game and corresponding expected utility of cooperation and defection. a Two-player public goods game, with the discrete options cooperation, C, and defection, D. It must hold that $b > c, b < 2c$. **b** Ego's expected utility for C and D with $b = 1.99$ and $c = 1$ in the public goods game, assuming self-regarding preferences. $E(C)$ exceeds $E(D)$ independently of the probability of other choosing C. **c** Assuming social preferences, the nominal public goods game transforms into a coordination game (averages of disadvantageous inequity aversion, α, and advantageous inequity aversion, β, used for calculations as reported in Fehr and Schmidt (1999)). **d** This transformation takes places when ego's subjective probability q that other will cooperate exceeds the threshold q^*. This threshold derives from equating $E(C)$ with $E(D)$ and solving for q^*, which yields $(2c(1+\alpha)-b)/(2c(\alpha+\beta))$. In this example, q^*equals to 0.6.

References

Abrahamse, W. and Steg, L. (2013) "Social influence approaches to encourage resource conservation: A meta-analysis," *Global Environmental Change*, 23(6), pp. 1773–1785. doi: 10.1016/j.gloenvcha.2013.07.029.

Ackermann, K. A. and Murphy, R. O. (2019) "Explaining cooperative behavior in public goods games: How preferences and beliefs affect contribution levels," *Games*, 10(1). doi: 10.3390/g10010015.

Allcott, H. and Rogers, T. (2014) "The short-run and long-run effects of behavioral interventions: Experimental evidence from energy conservation," *American Economic Review*, pp. 3003–3037. doi: 10.1257/aer.104.10.3003.

Amel, E. *et al.* (2017) "Beyond the roots of human inaction: Fostering collective effort toward ecosystem conservation," *Science*. American Association for the Advancement of Science, pp. 275–279. doi: 10.1126/science.aal1931.

Andreoni, J., Nikiforakis, N. and Siegenthaler, S. (2021) "Predicting social tipping and norm change in controlled experiments," *Proceedings of the National Academy of Sciences of the United States of America*, 118(16). doi: 10.1073/pnas.2014893118.

Battista, W. *et al.* (2018) "Behavior change interventions to reduce illegal fishing," *Frontiers in Marine Science*, 5(OCT). doi: 10.3389/fmars.2018.00403.

Berger, J. (2021) "Social tipping interventions can promote the diffusion or decay of sustainable consumption norms in the field. Evidence from a quasi-experimental intervention study," *Sustainability (Switzerland)*, 13(6), p. 3529. doi:

10.3390/su13063529.

Berger, J., Vogt, S. and Efferson, C. (2021) "Pre-existing fairness concerns restrict the cultural evolution and generalization of inequitable norms in children," *Evolution and Human Behavior*. Elsevier Inc., (July). doi: 10.1016/j.evolhumbehav.2021.07.001.

Bergquist, M. (2020) "Most People Think They Are More Pro-Environmental than Others: A Demonstration of the Better-than-Average Effect in Perceived Pro-Environmental Behavioral Engagement," *Basic and Applied Social Psychology*. Routledge, 42(1), pp. 50–61. doi: 10.1080/01973533.2019.1689364.

Bergquist, M., Nilsson, A. and Schultz, W. P. (2019) "A meta-analysis of field-experiments using social norms to promote pro-environmental behaviors," *Global Environmental Change*. Elsevier Ltd, 59(July), p. 101941. doi: 10.1016/j.gloenvcha.2019.101941.

Bicchieri, C. (2005) *The grammar of society: The nature and dynamics of social norms*, *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511616037.

Bicchieri, C. (2017) *Norms in the wild: How to diagnose, measure, and change social norms*, *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780190622046.001.0001.

Bicchieri, C. and Dimant, E. (2019) "Nudging with care: the risks and benefits of social information," *Public Choice*. Springer New York LLC. doi: 10.1007/s11127-019-00684-6.

Bicchieri, C. and Xiao, E. (2009) "Do the right thing: But only if others do so," *Journal of Behavioral Decision Making*. John Wiley and Sons Ltd, 22(2), pp. 191–208. doi: 10.1002/bdm.621.

Biel, A. and Thøgersen, J. (2007) "Activation of social norms in social dilemmas: A review of the evidence and reflections on the implications for environmental behaviour," *Journal of Economic Psychology*, 28(1), pp. 93–112. doi: 10.1016/j.joep.2006.03.003.

Blanco, M., Engelmann, D. and Normann, H. T. (2011) "A within-subject analysis of other-regarding preferences," *Games and Economic Behavior*, 72(2), pp. 321–338. doi: 10.1016/j.geb.2010.09.008.

Bluffstone, R. *et al.* (2020) "Cooperative behavior and common pool resources: Experimental evidence from community forest user groups in Nepal," *World Development*, 129(June). doi: 10.1016/j.worlddev.2020.104889.

Bowles, S. (2009) *Microeconomics: Behavior, institutions, and evolution*, *Microeconomics: Behavior, Institutions, and Evolution*. New York: Sage.

Boyd, R. *et al.* (2012) "Tragedy Revisited," *Science*, 47(42), pp. 14–17.

Brooks, J. S. *et al.* (2018) "Applying cultural evolution to sustainability challenges: an introduction to the special issue," *Sustainability Science*. doi: 10.1007/s11625-017-0516-3.

Byerly, H. *et al.* (2018) "Nudging pro-environmental behavior: evidence and opportunities," *Frontiers in Ecology and the Environment*, 16(3), pp. 159–168. doi: 10.1002/fee.1777.

Camerer, C. F. (2003) *Behavioral game theory: Experiments in strategic interaction*, *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton: Princeton University Press. doi: 10.1016/j.socec.2003.10.009.

Castilla-Rho, J. C. *et al.* (2017) "Social tipping points in global groundwater management," *Nature Human Behaviour*, 1(9), pp. 640–649. doi: 10.1038/s41562-017-0181-7.

Centola, D. *et al.* (2018) "Experimental evidence for tipping points in social convention," *Science*, 360(6393), pp. 1116–1119. doi: 10.1126/science.aas8827.

Centola, D. M. (2013) "Homophily, networks, and critical mass: Solving the start-up problem in large group collective action," *Rationality and Society*, 25(1), pp. 3–40. doi: 10.1177/1043463112473734.

Chaudhuri, A. (2011) "Sustaining cooperation in laboratory public goods experiments: A selective survey of the literature," *Experimental Economics*, 14(1), pp. 47–83. doi: 10.1007/s10683-010-9257-1.

Cinner, J. (2018) "How behavioral science can help conservation," *Science*, 362(6417), pp. 889–890. doi: 10.1126/science.aau6028.

Delmas, M. A., Fischlein, M. and Asensio, O. I. (2013) "Information strategies and energy conservation behavior: A meta-analysis of experimental studies from 1975 to 2012," *Energy Policy*, 61, pp. 729–739. doi: 10.1016/j.enpol.2013.05.109.

Diekmann, A. and Preisendörfer, P. (2003) "Green and greenback: The behavioral effects of environmental attitudes in low-cost and high-cost situations," *Rationality and Society*, 15(4), pp. 441–472. doi: 10.1177/1043463103154002.

Doell, K. C. *et al.* (2021) "Understanding the effects of partisan identity on climate change," *Current Opinion in Behavioral Sciences*, 42, pp. 54–59. doi: 10.1016/j.cobeha.2021.03.013.

Dolan, P. *et al.* (2012) "Influencing behaviour: The mindspace way," *Journal of Economic Psychology*, 33(1), pp. 264–277. doi: 10.1016/j.joep.2011.10.009.

Efferson, C. (2021) "Policy to activate cultural change to amplify policy," *Proceedings of the National Academy of Sciences of the United States of America*, 118(23), p.

e2106306118. doi: 10.1073/pnas.2106306118.

Efferson, C., Vogt, S. and Fehr, E. (2020) "The promise and the peril of using social influence to reverse harmful traditions," *Nature Human Behaviour*. Springer US, 4(1), pp. 55–68. doi: 10.1038/s41562-019-0768-2.

Engelmann, D. and Strobel, M. (2004) "Inequality aversion, efficiency, and maximin preferences in simple distribution experiments," *American Economic Review*, pp. 857–869. doi: 10.1257/0002828042002741.

Farjam, M., Nikolaychuk, O. and Bravo, G. (2019) "Experimental evidence of an environmental attitude-behavior gap in high-cost situations," *Ecological Economics*. Elsevier B.V., 166(166), p. 106434. doi: 10.1016/j.ecolecon.2019.106434.

Farrow, K., Grolleau, G. and Ibanez, L. (2017) "Social Norms and Pro-environmental Behavior: A Review of the Evidence," *Ecological Economics*, pp. 1–13. doi: 10.1016/j.ecolecon.2017.04.017.

Fehr, E. and Fischbacher, U. (2003) "The nature of human altruism," *Nature*, 425(6960), pp. 785–791. doi: 10.1038/nature02043.

Fehr, E. and Gächter, S. (2002) "Altruistic punishment in humans," *Nature*, 415(6868), pp. 137–140. doi: 10.1038/415137a.

Fehr, E. and Leibbrandt, A. (2011) "A field study on cooperativeness and impatience in the Tragedy of the Commons," *Journal of Public Economics*. Elsevier B.V., 95(9–10), pp. 1144–1155. doi: 10.1016/j.jpubeco.2011.05.013.

Fehr, E. and Schmidt, K. M. (1999) "A theory of fairness, competition, and cooperation," *Quarterly Journal of Economics*, 114(3), pp. 817–868. doi: 10.1162/003355399556151.

Fischbacher, U., Gächter, S. and Fehr, E. (2001) "Are people conditionally cooperative? Evidence from a public goods experiment," *Economics Letters*, 71(3), pp. 397–404. doi: 10.1016/S0165-1765(01)00394-9.

Frey, B. S. and Oberholzer-Gee, F. (1997) "The Cost of Price Incentives: An Empirical Analysis of Motivation Crowding-Out," *American Economic Review*, 87(4), pp. 746–755. doi: 10.2307/2951373.

Gintis, H. *et al.* (2003) "Explaining altruistic behavior in humans," *Evolution and Human Behavior*, 24(3), pp. 153–172. doi: 10.1016/S1090-5138(02)00157-5.

Goeschl, T. *et al.* (2020) "How much can we learn about voluntary climate action from behavior in public goods games?," *Ecological Economics*, 171. doi: 10.1016/j.ecolecon.2020.106591.

Goldstein, N. J., Cialdini, R. B. and Griskevicius, V. (2008) "A room with a viewpoint: Using social norms to motivate environmental conservation in hotels," *Journal of Consumer Research*, 35(3), pp. 472–482. doi: 10.1086/586910.

Gössling, S., Humpe, A. and Bausch, T. (2020) "Does 'flight shame' affect social norms? Changing perspectives on the desirability of air travel in Germany," *Journal of Cleaner Production*, 266. doi: 10.1016/j.jclepro.2020.122015.

Granovetter, M. (1978) "Threshold Models of Collective Behavior," *American Journal of Sociology*. University of Chicago Press, 83(6), pp. 1420–1443. doi: 10.1086/226707.

Guala, F. (2012) "Reciprocity: Weak or strong? What punishment experiments do (and do not) demonstrate," *Behavioral and Brain Sciences*, 35(1), pp. 1–15. doi: 10.1017/S0140525X11000069.

Hardin, G. (1968) "The tragedy of commons," *Science*, 162(December), pp. 1243–1248.

Hauser, O. P. *et al.* (2016) "Think global, act local: Preserving the global commons," *Scientific Reports*, 6. doi: 10.1038/srep36079.

Henrich, J. *et al.* (2001) "In search of Homo economicus: Behavioral experiments in 15 small-scale societies," *American Economic Review*, 91(2), pp. 73–84. doi: 10.1257/aer.91.2.73.

Henrich, J. *et al.* (2006) "Costly punishment across human societies," *Science*, 312(5781), pp. 1767–1770. doi: 10.1126/science.1127333.

Henrich, J. *et al.* (2010) "Markets, religion, community size, and the evolution of fairness and punishment," *Science*, 327(5972), pp. 1480–1484. doi: 10.1126/science.1182238.

Henrich, J., Heine, S. J. and Norenzayan, A. (2010a) "Most people are not WEIRD," *Nature*, p. 29. doi: 10.1038/466029a.

Henrich, J., Heine, S. J. and Norenzayan, A. (2010b) "The weirdest people in the world?," *Behavioral and Brain Sciences*, pp. 61–83. doi: 10.1017/S0140525X0999152X.

Irwin, T. (2009) *Implications for climate-change policy of research on cooperation in social dilemmas*, *World Bank Policy Research Working Paper Series*. Washington DC. Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1437958.

Kahan, D. (2012) "Why we are poles apart on climate change," *Nature*, p. 255. doi: 10.1038/488255a.

Kahan, D. M. (2017) "The expressive rationality of inaccurate perceptions," *Behavioral and Brain Sciences*. doi: 10.1017/S0140525X15002332.

Kerr, J. M. *et al.* (2017) "Long-term effects of payments for environmental services: Combining insights from communication and economics," *Sustainability*

(Switzerland). MDPI AG, 9(9). doi: 10.3390/su9091627.

Kerr, J. M. *et al.* (2019) "The effects of social norms on motivation crowding: Experimental evidence from the tibetan plateau," *International Journal of the Commons*, 13(1), pp. 430–454. doi: 10.18352/ijc.882.

Kerr, J., Vardhan, M. and Jindal, R. (2012) "Prosocial behavior and incentives: Evidence from field experiments in rural Mexico and Tanzania," *Ecological Economics*. Elsevier B.V., 73, pp. 220–227. doi: 10.1016/j.ecolecon.2011.10.031.

Klotz, L. *et al.* (2018) "Beyond rationality in engineering design for sustainability," *Nature Sustainability*. Springer US, 1(5), pp. 225–233. doi: 10.1038/s41893-018-0054-8.

Lapinski, M. K. *et al.* (2017) "Social Norms, Behavioral Payment Programs, and Cooperative Behaviors: Toward a Theory of Financial Incentives in Normative Systems," *Human Communication Research*, 43(1), pp. 148–171. doi: 10.1111/hcre.12099.

Lee-Rife, S. *et al.* (2012) "What Works to Prevent Child Marriage: A Review of the Evidence," *Studies in Family Planning*, 43(4), pp. 287–303. doi: 10.1111/j.1728-4465.2012.00327.x.

Leviston, Z. and Uren, H. V. (2020) "Overestimating One's 'Green' Behavior: Better-Than-Average Bias May Function to Reduce Perceived Personal Threat from Climate Change," *Journal of Social Issues*, 76(1), pp. 70–85. doi: 10.1111/josi.12365.

Liebe, U., Gewinner, J. and Diekmann, A. (2021) "Large and persistent effects of green energy defaults in the household and business sectors," *Nature Human Behaviour*. Springer US, 5(5), pp. 576–585. doi: 10.1038/s41562-021-01070-3.

Lokhorst, A. M. *et al.* (2013) "Commitment and Behavior Change: A Meta-Analysis and

Critical Review of Commitment-Making Strategies in Environmental Research," *Environment and Behavior*, 45(1), pp. 3–34. doi: 10.1177/0013916511411477.

Mackie, G. (1996) "Ending footbinding and infibulation: A convention account," *American Sociological Review*. [American Sociological Association, Sage Publications, Inc.], 61(6), pp. 999–1017. doi: 10.2307/2096305.

Mackie, G. and Lejeune, J. (2009) "Social Dynamics of Abandonment of Harmful Practices: a New Look at the Theory," *UNICEF Innocenti Working paper*, (May).

Macy, M. W. (1991) "Chains of Cooperation: Threshold Effects in Collective Action," *American Sociological Review*, 56(6), p. 730. doi: 10.2307/2096252.

Mathias, J. D. *et al.* (2020) "Exploring non-linear transition pathways in social-ecological systems," *Scientific Reports*, 10(1). doi: 10.1038/s41598-020-59713-w.

Mccright, A. M. and Dunlap, R. E. (2011) "The Politicization Of Climate Change And Polarization In The American Public's Views Of Global Warming, 2001-2010," *Sociological Quarterly*, 52(2), pp. 155–194. doi: 10.1111/j.1533-8525.2011.01198.x.

Milinski, M., Semmann, D. and Krambeck, H. J. (2002) "Reputation helps solve the 'tragedy of the commons,'" *Nature*, 415(6870), pp. 424–426. doi: 10.1038/415424a.

Milkoreit, M. *et al.* (2018) "Defining tipping points for social-ecological systems scholarship - An interdisciplinary literature review," *Environmental Research Letters*. doi: 10.1088/1748-9326/aaaa75.

Moser, S. and Kleinhückelkotten, S. (2018) "Good Intents, but Low Impacts: Diverging Importance of Motivational and Socioeconomic Determinants Explaining Pro-Environmental Behavior, Energy Use, and Carbon Footprint," *Environment and Behavior*, 50(6), pp. 626–656. doi: 10.1177/0013916517710685.

Muthukrishna, M. and Schaller, M. (2020) "Are Collectivistic Cultures More Prone to Rapid Transformation? Computational Models of Cross-Cultural Differences, Social Network Structure, Dynamic Social Influence, and Cultural Change," *Personality and Social Psychology Review*, 24(2), pp. 103–120. doi: 10.1177/1088868319855783.

Nisa, C. F. *et al.* (2019) "Meta-analysis of randomised controlled trials testing behavioural interventions to promote household action on climate change," *Nature Communications*, 10(1). doi: 10.1038/s41467-019-12457-2.

Nyborg, K. *et al.* (2016) "Social norms as solutions," *Science*, pp. 42–43. doi: 10.1126/science.aaf8317.

Nyborg, K. (2018) "Social Norms and the Environment," *Annual Review of Resource Economics*, 10, pp. 405–423. doi: 10.1146/annurev-resource-100517-023232.

Nyborg, K. (2020) "No Man is an Island: Social Coordination and the Environment," *Environmental and Resource Economics*, 76(1), pp. 177–193. doi: 10.1007/s10640-020-00415-2.

Nyborg, K. and Rege, M. (2003) "On social norms: The evolution of considerate smoking behavior," *Journal of Economic Behavior and Organization*, 52(3), pp. 323–340. doi: 10.1016/S0167-2681(03)00031-3.

O'Brien, C. (2017) "When Norms Collide: Local Responses to Activism against Female Genital Mutilation and Early Marriage, Karisa Cloward (New York: Oxford University Press, 2016), 332 pp., $99 cloth, $34.95 paper.," *Ethics & International Affairs*, 31(3), pp. 388–390. doi: 10.1017/s0892679417000284.

Organisation, W. H. (2009) "Changing cultural and social norms that support violence Series of briefings on violence prevention," *Violence Injury Prevention, World Health*

Organization.

Osbaldiston, R. and Schott, J. P. (2012) "Environmental sustainability and behavioral science: Meta-analysis of proenvironmental behavior experiments," *Environment and Behavior*, 44(2), pp. 257–299. doi: 10.1177/0013916511402673.

Ostrom, E. (2015) *Governing the commons: The evolution of institutions for collective action*, *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9781316423936.

Otto, I. M. *et al.* (2020) "Social tipping dynamics for stabilizing Earth's climate by 2050," *Proceedings of the National Academy of Sciences of the United States of America*, 117(5), pp. 2354–2365. doi: 10.1073/pnas.1900577117.

Pieters, R. *et al.* (1998) "Consumers' attributions of proenvironmental behavior, motivation, and ability to self and others," *Journal of Public Policy and Marketing*, 17(2), pp. 215–225. doi: 10.1177/074391569801700206.

Platteau, J. P. and Auriol, E. (2018) "Eradicating women-hurting customs," in Anderson, S., Beaman, L., and Plateau, J. (eds.) *Towards Gender Equity in Development*. Oxford University Press, pp. 319–356.

Reddy, S. M. W. *et al.* (2017) "Advancing Conservation by Understanding and Influencing Human Behavior," *Conservation Letters*, 10(2), pp. 248–256. doi: 10.1111/conl.12252.

Reeson, A. F. and Tisdell, J. G. (2008) "Institutions, motivations and public goods: An experimental test of motivational crowding," *Journal of Economic Behavior and Organization*, 68(1), pp. 273–281. doi: 10.1016/j.jebo.2008.04.002.

Rinscheid, A., Pianta, S. and Weber, E. U. (2020) "What shapes public support for climate

change mitigation policies? The role of descriptive social norms and elite cues," *Behavioural Public Policy*, pp. 1–25. doi: 10.1017/bpp.2020.43.

Rode, J., Gómez-Baggethun, E. and Krause, T. (2015) "Motivation crowding by economic incentives in conservation policy: A review of the empirical evidence," *Ecological Economics*, 117, pp. 270–282. doi: 10.1016/j.ecolecon.2014.11.019.

Rustagi, D., Stefanie, E. and Kosfeld, M. (2010) "Conditional cooperation and costly monitoring explain success in forest commons management," *Science*, 330(6006), pp. 961–965. doi: 10.1126/science.1193649.

Scheffer, M. (2020) *Critical Transitions in Nature and Society*, *Critical Transitions in Nature and Society*. Princeton: Princeton University Press. doi: 10.2307/j.ctv173f1g1.

Schelling, T. C. (1973) "Hockey Helmets, Concealed Weapons, and Daylight Saving: A Study of Binary Choices With Externalities," *Journal of Conflict Resolution*, 17(3), pp. 381–428. doi: 10.1177/002200277301700302.

Schelling, T. C. (1978) *Micromotives and macrobehavior*. New York: Norton.

Schiermeier, Q. (2019) "Eat less meat: UN climate-change report calls for change to human diet," *Nature*, pp. 291–292. doi: 10.1038/d41586-019-02409-7.

Schultz, P. W. *et al.* (2007) "The constructive, destructive, and reconstructive power of social norms: Research article," *Psychological Science*, 18(5), pp. 429–434. doi: 10.1111/j.1467-9280.2007.01917.x.

Shell-Duncan, B. *et al.* (2011) "Dynamics of change in the practice of female genital cutting in Senegambia: Testing predictions of social convention theory," *Social Science and Medicine*, 73(8), pp. 1275–1283. doi: 10.1016/j.socscimed.2011.07.022.

Sigdel, R. P., Anand, M. and Bauch, C. T. (2017) "Competition between injunctive social

norms and conservation priorities gives rise to complex dynamics in a model of forest growth and opinion dynamics," *Journal of Theoretical Biology*. Elsevier Ltd, 432, pp. 132–140. doi: 10.1016/j.jtbi.2017.07.029.

Steg, L. *et al.* (2014) "An Integrated Framework for Encouraging Pro-environmental Behaviour: The role of values, situational factors and goals," *Journal of Environmental Psychology*, pp. 104–115. doi: 10.1016/j.jenvp.2014.01.002.

Sugden, R., Harsanyi, J. C. and Selten, R. (1989) *A General Theory of Equilibrium Selection in Games.*, *Economica*. Cambridge, MA: MIT Press. doi: 10.2307/2554329.

Thaler, R. H. and Sunstein, C. R. (2008) *Nudge: Improving decisions about health, wealth, and happiness*, *Nudge: Improving Decisions about Health, Wealth, and Happiness*. New Haven: Yale University Press. doi: 10.1016/s1477-3880(15)30073-6.

Travers, H. *et al.* (2021) "Delivering behavioural change at scale: What conservation can learn from other fields," *Biological Conservation*. doi: 10.1016/j.biocon.2021.109092.

Verkooijen, K. T., Stok, F. M. and Mollen, S. (2015) "The power of regression to the mean: A social norm study revisited," *European Journal of Social Psychology*, 45(4). doi: 10.1002/ejsp.2111.

Walker, B. and Meyers, J. A. (2004) "Thresholds in ecological and social-ecological systems: A developing database," *Ecology and Society*, 9(2). doi: 10.5751/es-00664-090203.

Waring, T. M. *et al.* (2015) "A multilevel evolutionary framework for sustainability analysis," *Ecology and Society*, 20(2). doi: 10.5751/ES-07634-200234.

Waring, T. M., Goff, S. H. and Smaldino, P. E. (2017) "The coevolution of economic institutions and sustainable consumption via cultural group selection," *Ecological*

Economics, 131, pp. 524–532. doi: 10.1016/j.ecolecon.2016.09.022.

Westley, F. *et al.* (2011) "Tipping toward sustainability: Emerging pathways of transformation," in *Ambio*, pp. 762–780. doi: 10.1007/s13280-011-0186-9.

World Bank Group (2015) *Mind, Society, and Behavior: World Development Report*.

Wynes, S. and Nicholas, K. A. (2017) "The climate mitigation gap: Education and government recommendations miss the most effective individual actions," *Environmental Research Letters*, 12(7). doi: 10.1088/1748-9326/aa7541.

Young, H. P. (2015) "The Evolution of Social Norms," *Annual Review of Economics*, 7(1), pp. 359–387. doi: 10.1146/annurev-economics-080614-115322.

Figures

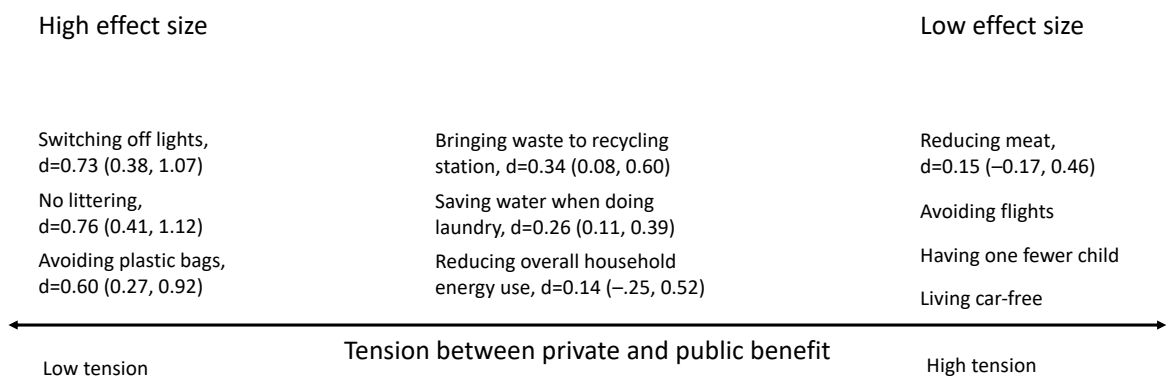


Figure 1. Effect of a behaviour intervention on pro-environmental behaviour, depending on the tension between the private and public benefit inherent to the underlying environmental dilemma. Low tension implies room for strong effects. High tension implies little room for strong effects. Effect sizes are taken from Bergquist et al., (2019) and Nisa et al., (2019).

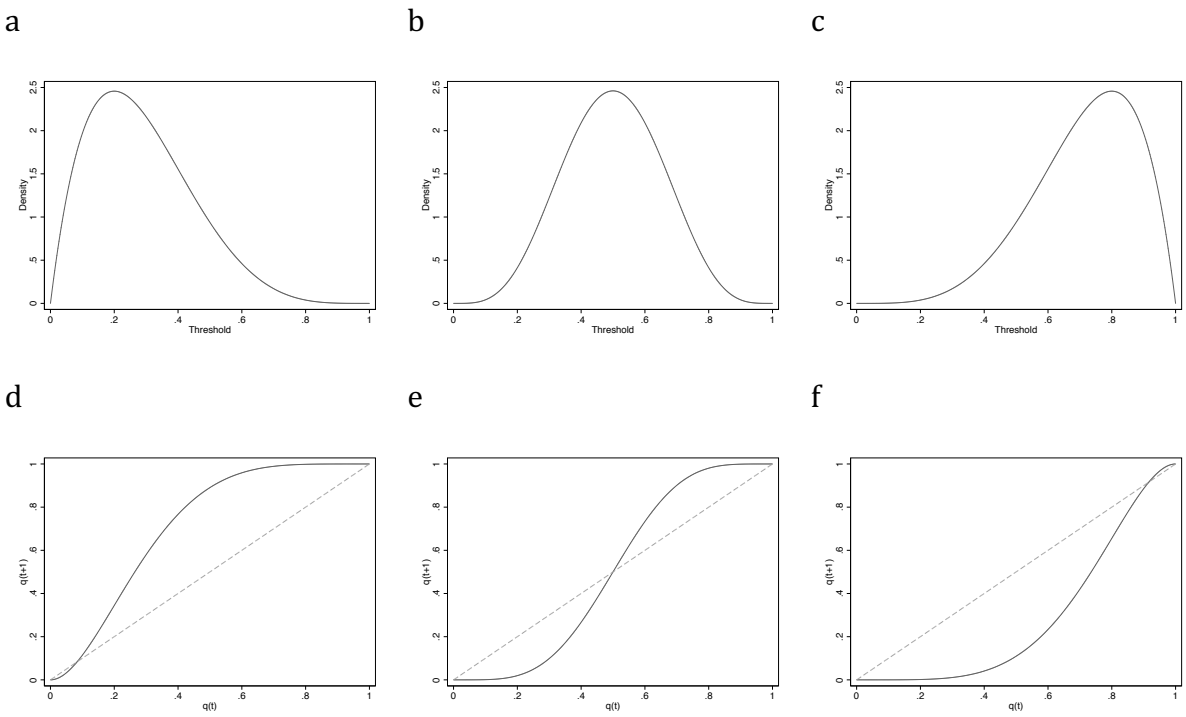


Figure 2. Illustrative threshold density distributions (a-c), and the corresponding cumulative distributions (d-f). The shape of a given distribution depends on the tension between the private and public benefit inherent to the underlying environmental dilemma. **a, d** Low tension, right-skewed distribution. **b, e** Medium tension, symmetric distribution. **c, f** High tension, left-skewed distribution.

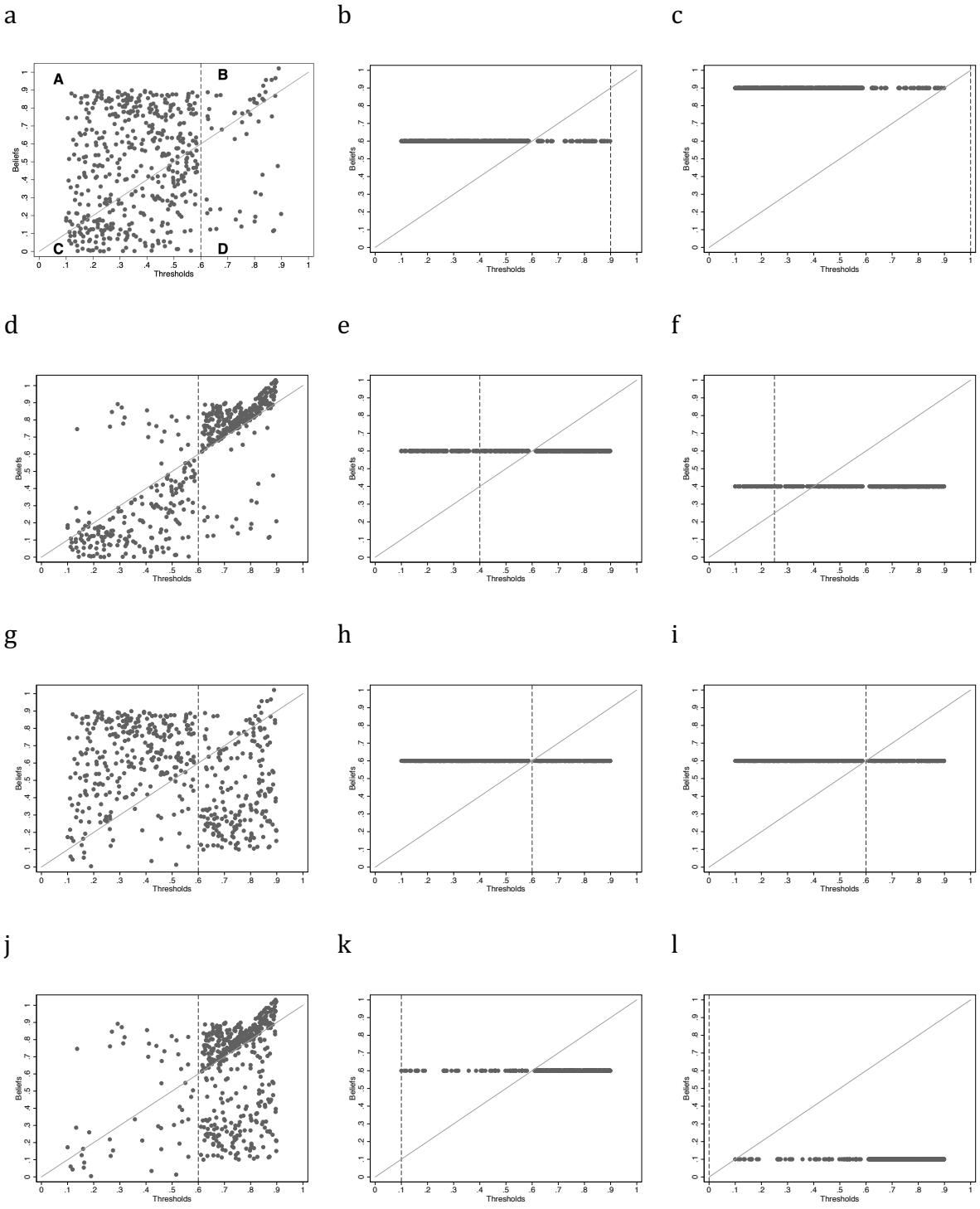


Figure 3. Beliefs-based intervention. 500 thresholds, q_i^*, and beliefs, $\hat{q}_{it}$, pre-intervention ($t = 0$) (**a, d, g, j**), after a first intervention ($t = 1$) (**b, e, h, k**) and after a second intervention ($t = 2$) (**c, f, I, l**). The dashed vertical line denotes cooperation q at time t. Region A: cooperators x_1 with the accurate belief of $\hat{q}_{i0} \geq q_i^*$. B: cooperators x_2 with the false belief of $\hat{q}_{i0} \geq q_i^*$. C: defectors x_3 with the false

belief of $\hat{q}_{i0} < q_i^*$. D: defectors x_4 with the accurate belief of $\hat{q}_{i0} < q_i^*$. Post-intervention ($t = 1$), all individuals adapt their beliefs to match the descriptive feedback. Concerning behaviour, the individuals in A keep cooperating, those in B switch from cooperation to defection, those in C start cooperating and those in D keep defecting. The same logic applies to a second intervention at $t = 2$, with post-intervention cooperation q_1 as the new baseline for decisions regarding cooperation. **a-c** Scenario 1: right-skewed distribution, $\hat{q}_{i0}$ and q_i^* uncorrelated. **d-f**. Scenario 2: symmetric distribution, $\hat{q}_{i0}$and q_i^* positively correlated. **g-i**. Scenario 3: symmetric distribution, $\hat{q}_{i0}$ and q_i^* negatively correlated. **j-l**. Scenario 4: left-skewed distribution, $\hat{q}_{i0}$ and q_i^* uncorrelated.

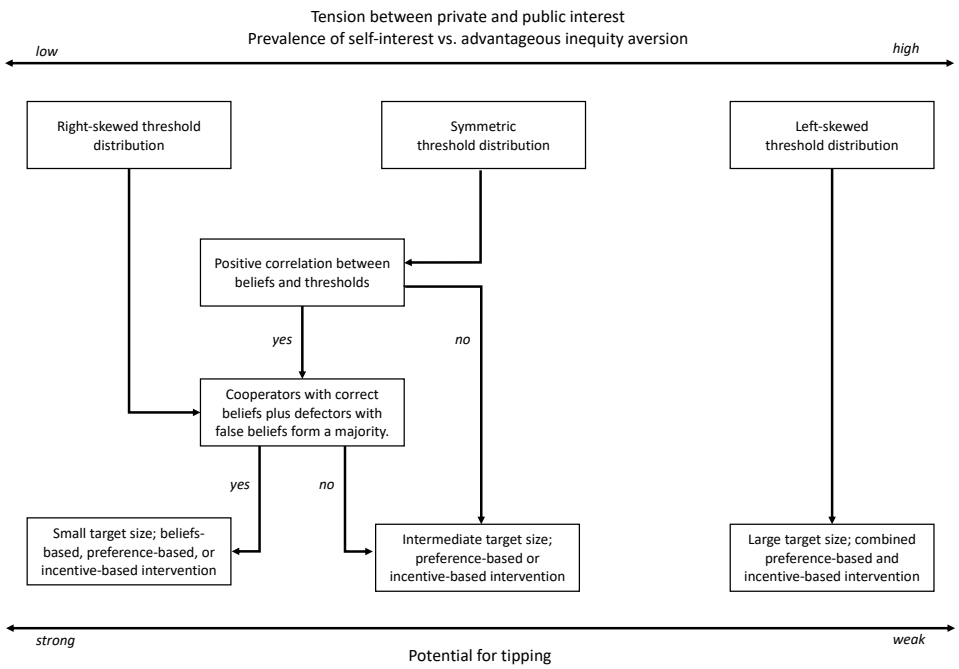


Figure 4. Structure of preference heterogeneity, the potential for tipping, and suggested intervention strategy. The tension between private and public benefit underlying an environmental dilemma, the distribution of pro-social preferences in a population, the distribution of beliefs regarding cooperation (pre-intervention) and the cooperation level (pre-intervention) jointly shape the potential for tipping. Small interventions, centring on the mere provision of descriptive feedback may activate tipping if the potential is weak. Large interventions, combining different intervention strategies are necessary to activate tipping if the potential is strong.